RNA-Seq workflow

Phase 2: Generating expression count data

Andrew Ndhlovu

Project SeaStore
Department of Botany and Zoology
Stellenbosch University
https://github.com/PiscatorX/RNA-Seq-devs

August 1, 2023

Overview of RNA-Seq workflow

Phase I

Check original (raw) data quality

Pre-process reads

- Trim adapter remnants
- Trim low quality bases (Phred score ≤ 25)
- Remove reads ≤ 20nt

Recheck data quality

Phase II

Generate gene/transcript level counts

- Align reads to reference genome, or
- Generate estimated counts using pseudo-alignment approaches

Tabulate overall summary statistics (depends on method used above)

Phase III

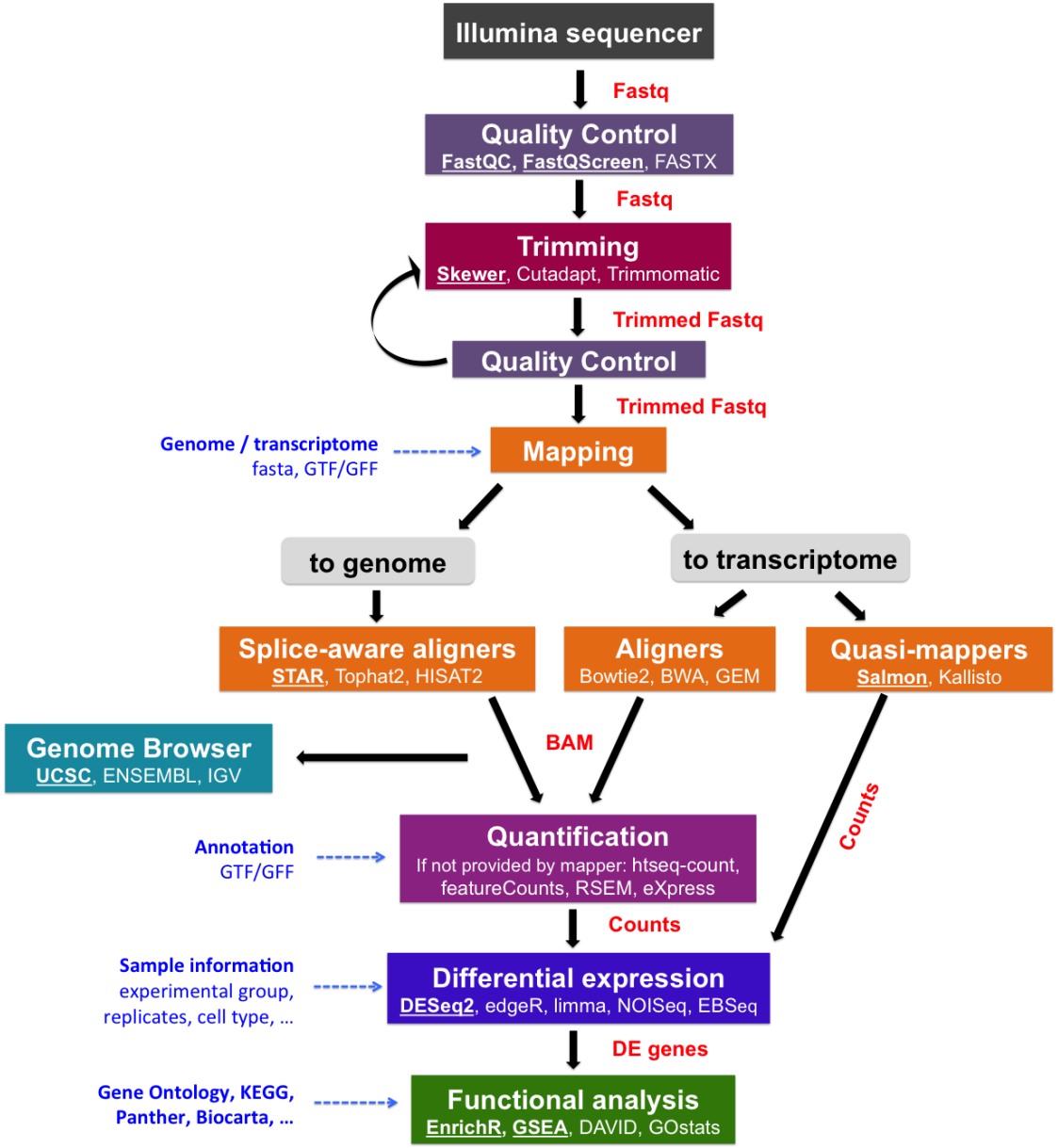
Perform quality checks on count data

- Check for outliers among replicates from same set
- Filter out any genes with counts below selected threshold

Perform statistical analysis to find differentially expressed genes

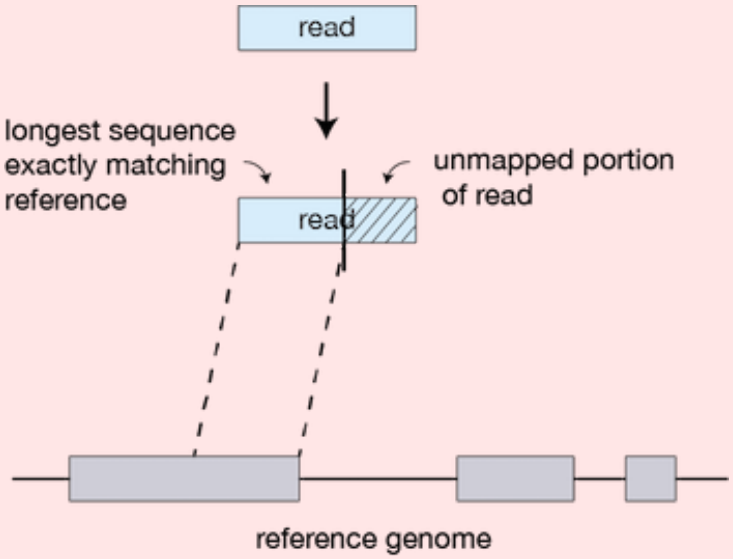
https://h3abionet.github.io/H3ABionet-SOPs/RNA-Seq-1-2.html

RNA-Seq goal is differential expression analysis



RNA-Seq alignment is key to expression quantification

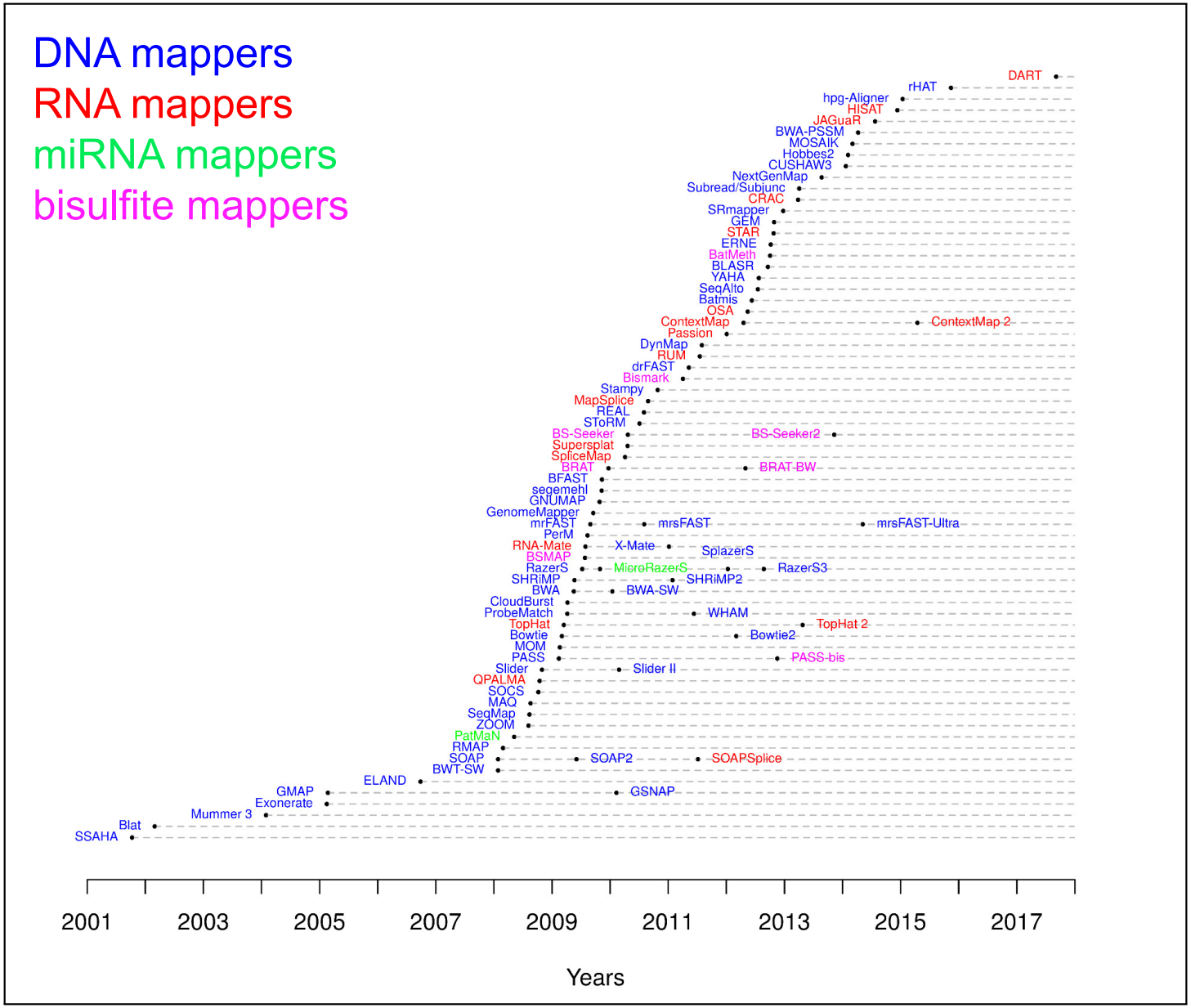
Challenges of aligning RNA-seq reads to a genome



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

- RNA reads do not contain introns (spliced out during transcription)
- Sequences may span multiple exons
- Mapping to a reference genome is computationally challenging
- Too many reads: millions to align!!!
- Need for splice aware aligners

Spoilt for choice: Lots of mappers



http://genomeast.igbmc.fr/wiki/lib/exe/fetch.php?media=training:190115_rnaseq_phdpgr.pdf

Lots of comparison studies

RESEARCH Ope

Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods

Brian J. Haas[1*], Alexander Dobin[2], Bo Li[1,3], Nicolas Stransky[4], Nathalie Pochet[1,5] and Aviv Regev[1,6]

ORIGINAL ARTICLE

New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies

Luigi Donato[1,2] · Concetta Scimone[1,2] · Carmela Rinaldi[1] · Rosalia D'Angelo[1] · Antonina Sidoti[1]

Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider

Ryan Musich[1], Lance Cadle-Davidson[1,2] and Michael V. Osier[1]*

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, United States, [2]USDA-Agricultural Research Service, Grape Genetics Research Unit, Geneva, NY, United States

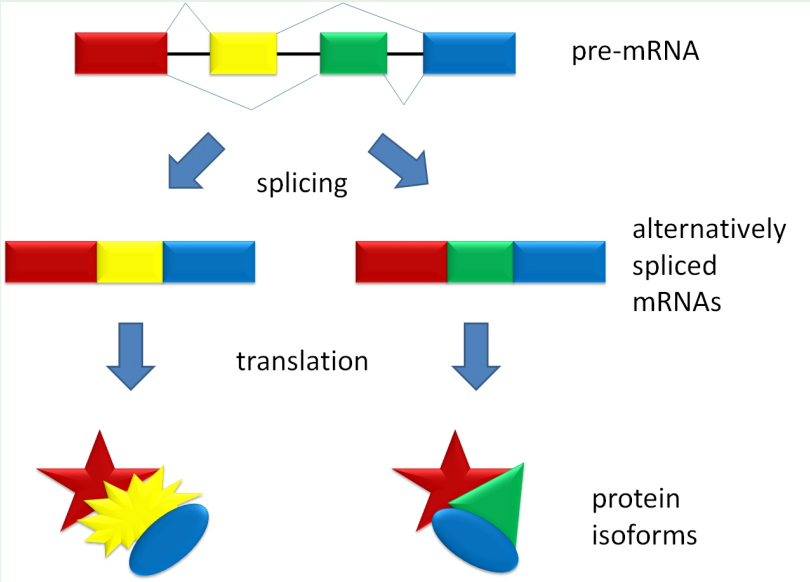
REVIEW Open Access

Technology dictates algorithms: recent developments in read alignment

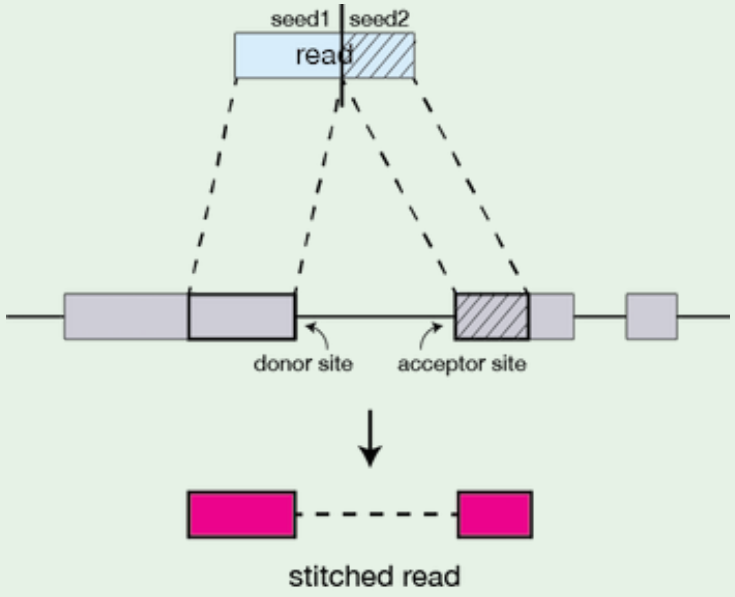
Mohammed Alser[1,2,3†], Jeremy Rotman[4†], Dhrithi Deshpande[5], Kodi Taraszka[4], Huwenbo Shi[6,7], Pelin Icer Baykal[8], Harry Taegyun Yang[4,9], Victor Xue[4], Sergey Knyazev[8], Benjamin D. Singer[10,11,12], Brunilda Balliu[13], David Koslicki[14,15,16], Pavel Skums[8], Alex Zelikovsky[8,17], Can Alkan[2,18], Onur Mutlu[1,2,3†] and Serghei Mangul[5*†]

STAR: Spliced Transcripts Alignment to a Reference

A fast splice aware aligner



https://simple.wikipedia.org/wiki/Alternative_splicing

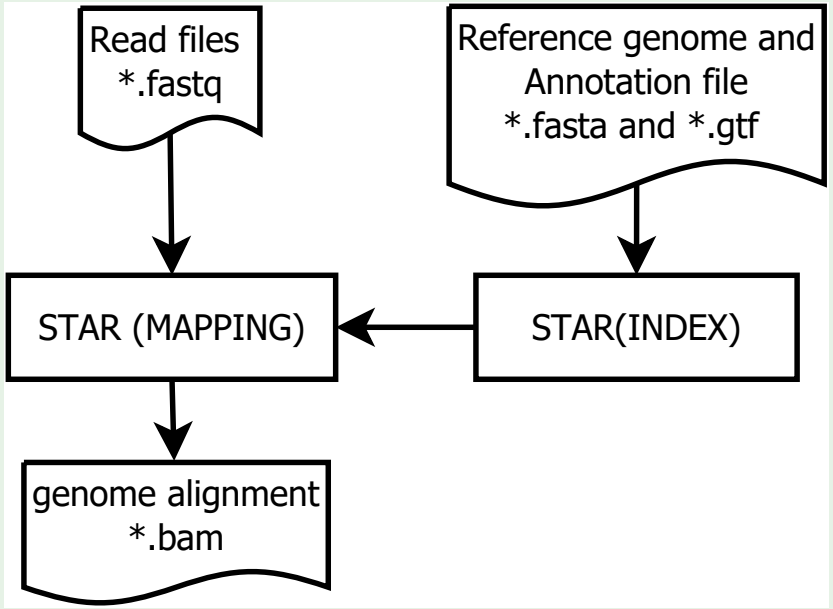


https://www.reneshbedre.com/blog/star-aligner.html

- STAR outperforms other aligners by $>$ 50X in mapping speed
- Discovers non-canonical splices and chimeric (fusion) transcripts
  - Aligns reads with indels due to genomic variations or sequencing errors.
  - Identifies spliced RNAs formed by sequences across genomic regions

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

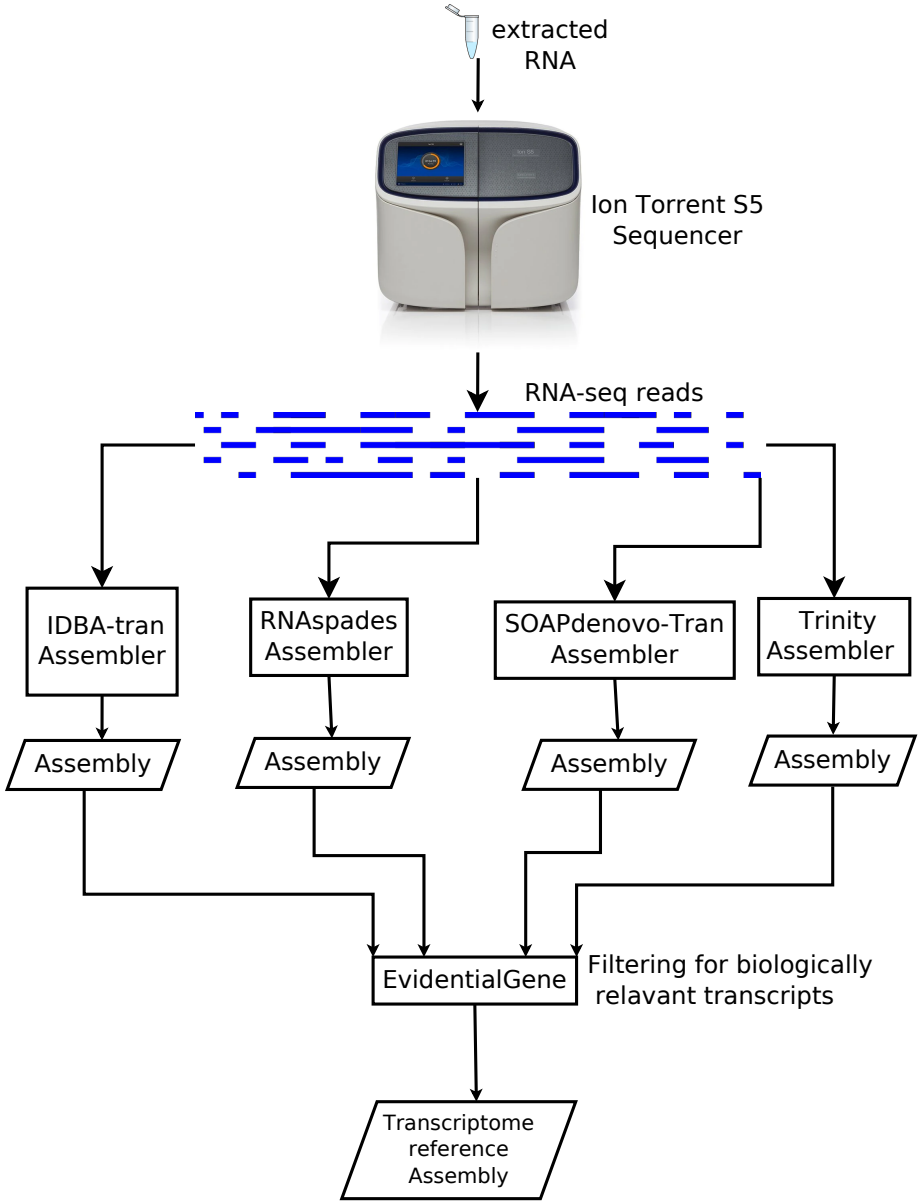
Basic STAR workflow consists of 2 steps



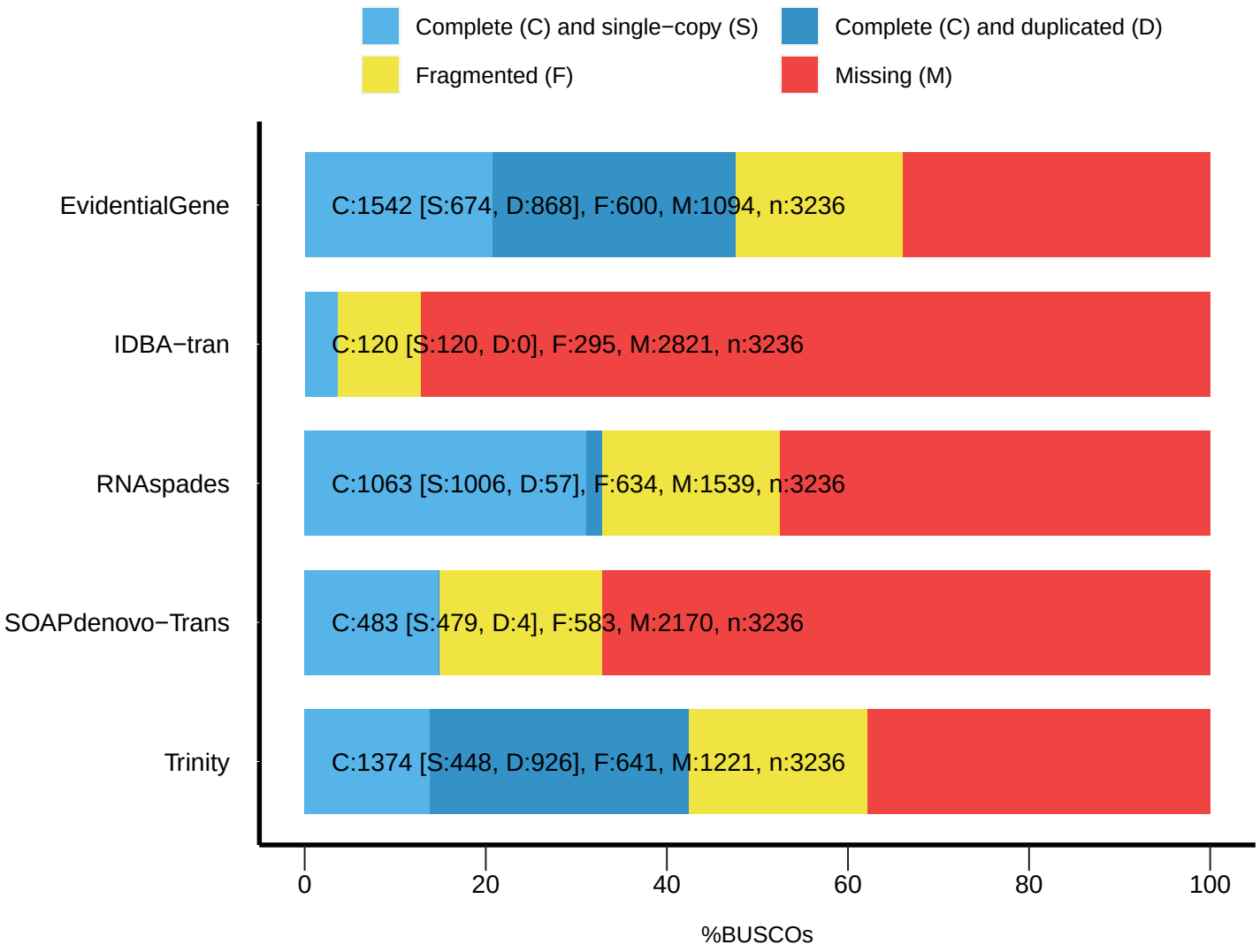
1. Generating genome indexes files.
   Reference genome sequences (FASTA files) and annotations (GTF file) to generate genome indexes
2. Mapping reads to the genome.
   STAR maps the reads to the genome index, and writes several output files, such as alignments (SAM/BAM), mapping summary statistics, splice junctions, unmapped reads, signal (wiggle) tracks etc.

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

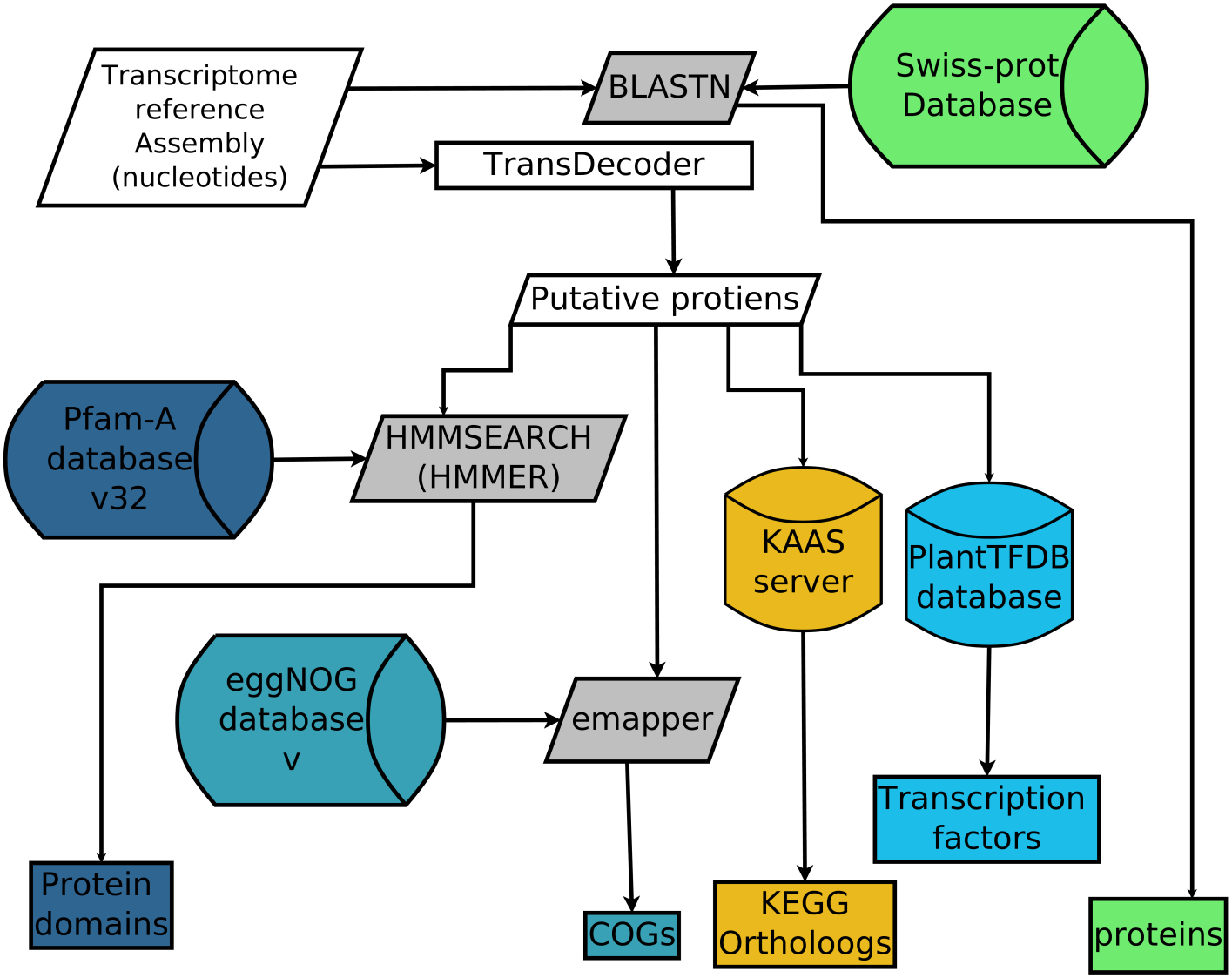
Sequencing and *denovo* transcriptome assembly



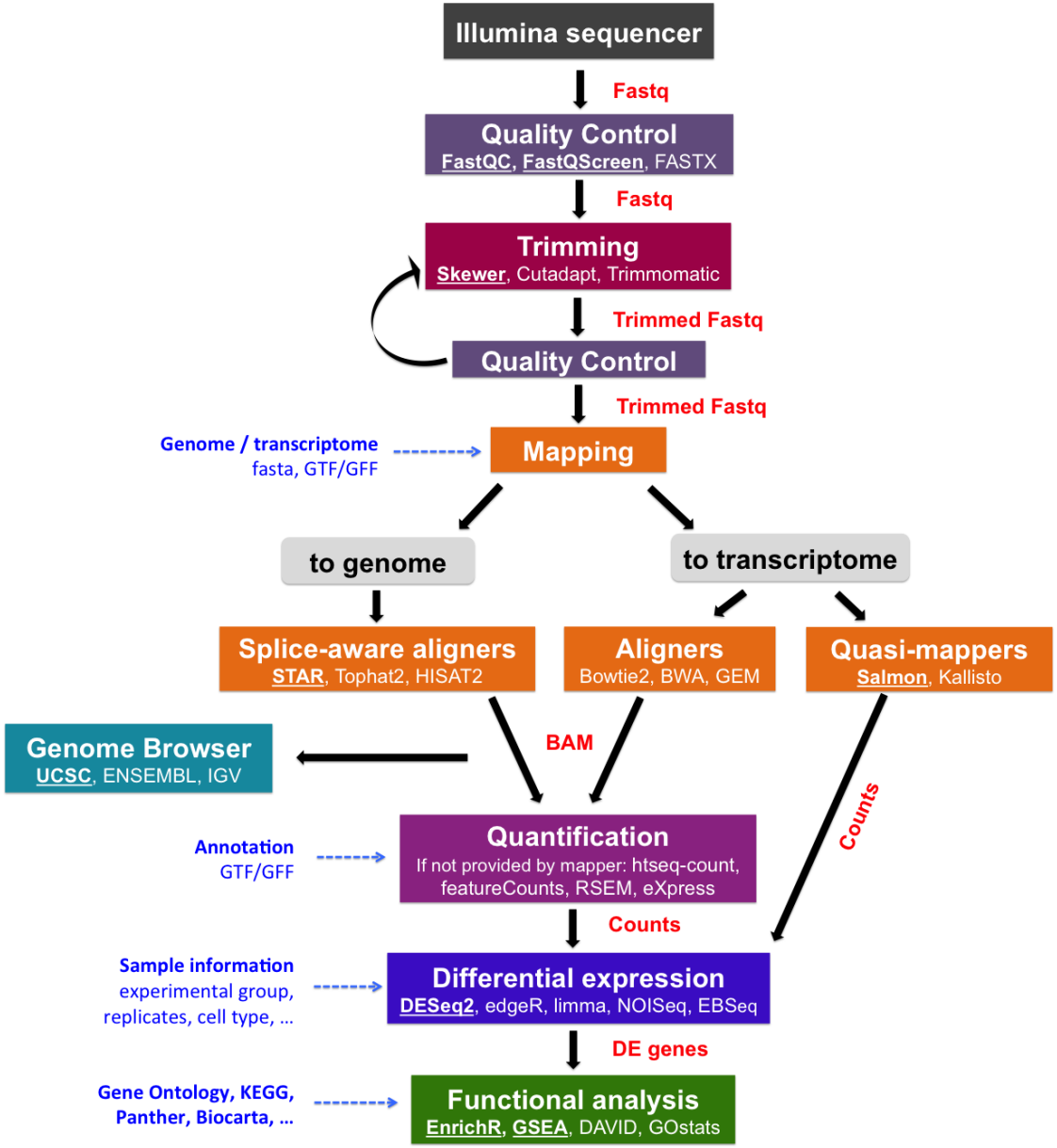
Benchmarking Universal Single-Copy Orthologs (BUSCOs)



Annotation of transcripts and predicted proteins

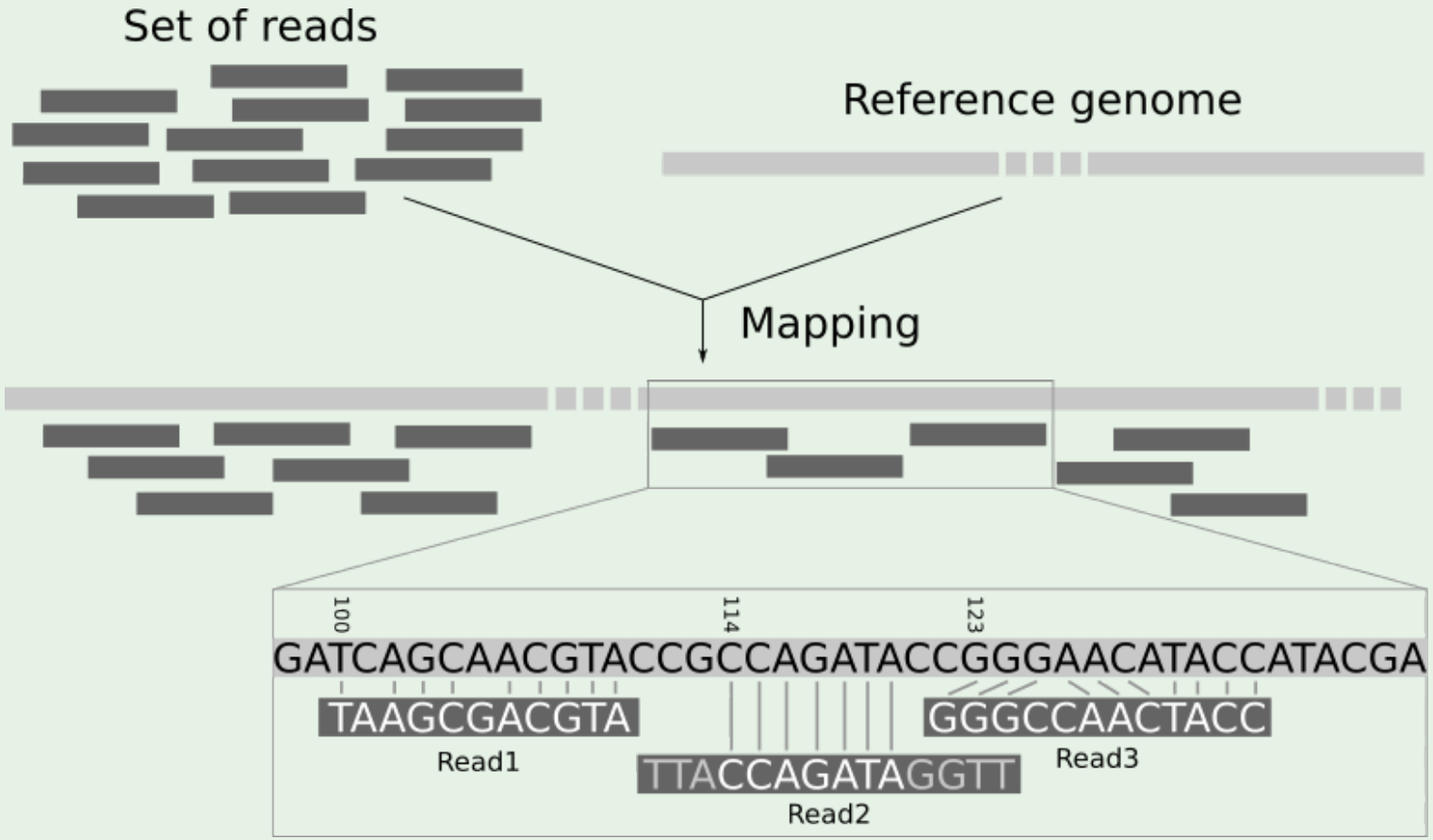


Recap: Goal of RNA-Seq is differential expression analysis



Mapping reads to the transcriptome reference

Bowtie 2: an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences



https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Read mapping workflow

Bowtie

https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

- Building an index from the transcriptome reference
- Aligning reads to reference. Output is a sequence alignment/map (SAM) file

samtools

http://www.htslib.org/

- Convert SAM to BAM (binary alignment map) file [samtools view -b <sam>]
- Sort BAM files using referenc [samtools sort <bam>]
- Index BAM [samtools index <bam>]
- Get alignment statistics [samtools flagstat <bam>]

What is in a SAM file

https://samtools.github.io/hts-specs/SAMv1.pdf

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       * 0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       * 0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         = 7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

CIGAR: Concise Idiosyncratic Gapped Alignment Report

https://www.samformat.info/sam-format-flag

Inspecting the bowtie mapping

Check samtools stats

https://davetang.org/wiki/tiki-index.php?page=SAMTools

```
samtools flagstat <bam>
```

Visualiase mapping using IGV

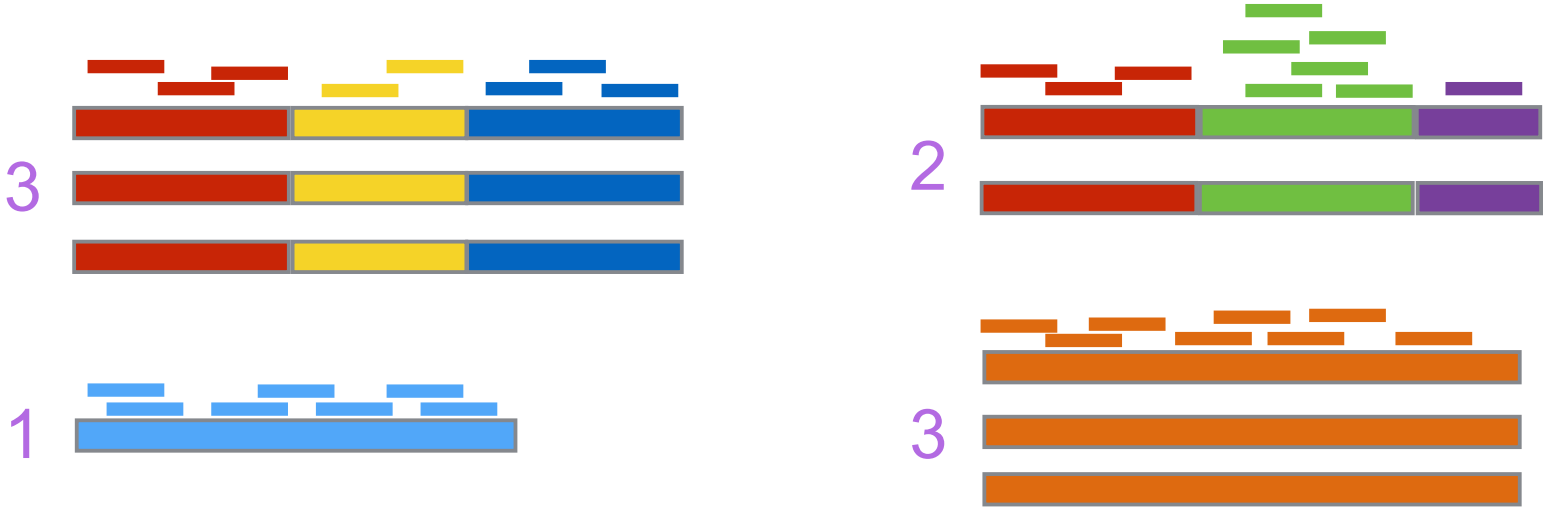
https://software.broadinstitute.org/software/igv/ Integrative Genomics Viewer (IGV) is an interactive tool for the visual exploration of genomic data.

- bam file
- bam file index
- reference sequence

Transcript/gene ~~quantification~~ estimation

Mapping of reads counts read but we want to count transcripts

Goal: estimate the abundance of each kind of transcript given short reads sampled from the expressed transcripts.



https://simons.berkeley.edu/sites/default/files/docs/4546/kingsford-regulatory-genomics-salmon.pdf

Challenges of transcript/gene estimation

Goal: estimate the abundance of each kind of transcript given short reads sampled from the expressed transcripts.

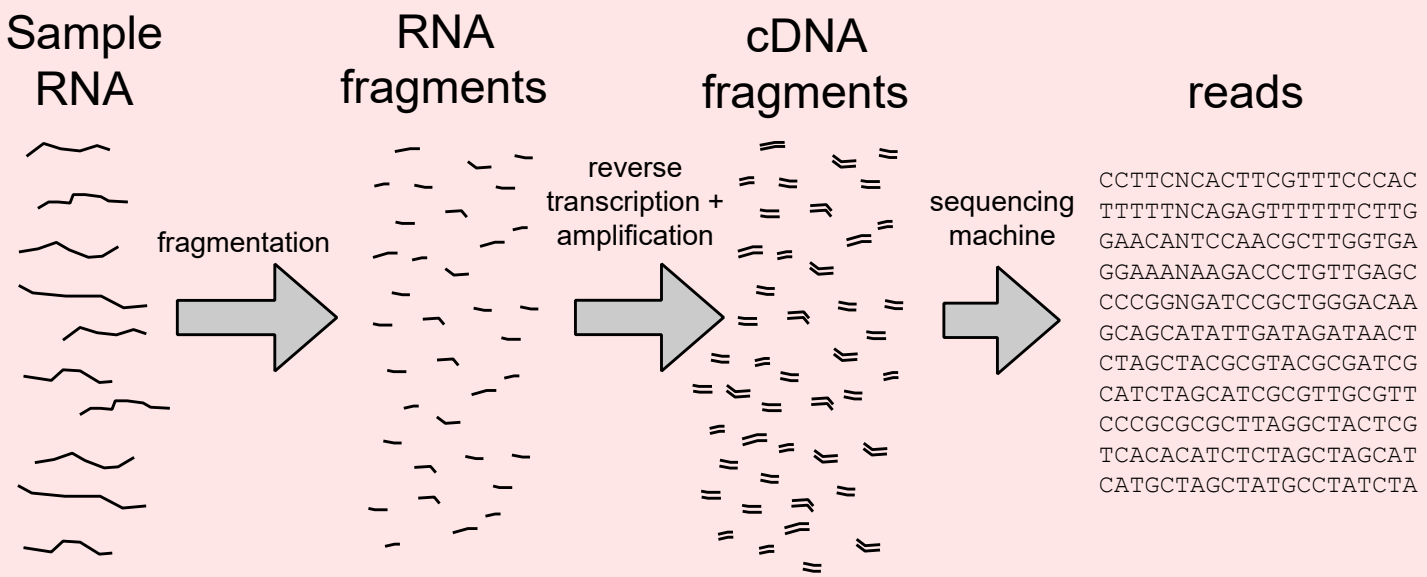
Challenges:

- Hundreds of millions of short reads per experiment
- Finding locations of reads (mapping) is traditionally slow
- Alternative splicing creates ambiguity about where reads came from
- Sampling of reads is not uniform
- How do we account for multiply mapped reads
- What about mapping quality
- How we account for paired end reads vs single end reads
- What about transcripts of different lengths
- How do we move from transcript to gene annotation data
- Models for counting the transcripts

https://simons.berkeley.edu/sites/default/files/docs/4546/kingsford-regulatory-genomics-salmon.pdf

RNA-Seq is a relative abundance measurement technology

The is a need for normalised data



- Reads are sequenced from fragments of transcripts
- We assume that reads are a random sample of fragments
- Absolute counts are not very useful as they are linked to library preparation and sequencing platform

Relative abundance is useful but not sufficient

Need to normalise our data

Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

- Changes in absolute expression of high expressors is a major factor
- Normalization is required for comparing samples in these situations

- This does not account for transcript length and multi-mapped reads
- You cannot compare across samples

RPKM: Reads Per Kilobase Million

https://www.novogene.com/us-en/resources/blog/how-to-choose-normalization-methods-tpm-rpkm-fpkm-for-mrna-expression/

$$RPKM = \frac{numReads}{\frac{geneLength}{1000} * \frac{totalNumReads}{1000000}}$$

$numReads$ = number of reads mapped to a gene sequence (SE end reads only)
$geneLength$ = Length of the gene or transcript sequence
$totalNumReads$ = total number of mapped reads of reads of a sample

FPKM: Fragments Reads Per Kilobase Million

$$FPKM = \frac{numReads}{\frac{geneLength}{1000} * \frac{totalNumReads}{1000000}}$$

$numReads$ = number of reads mapped to a gene sequence
(avoids double counting of reads mapped to the same transcript)
$geneLength$ = Length of the gene or transcript sequence
$totalNumReads$ = total number of mapped reads of reads of a sample

TPM: Transcripts Per kilobase Million

Introduced in an attempt to facilitate comparisons across samples

$$TPM = \frac{Ni/Li * 10^6}{sum(N1/L1 + N2/L2 + ... + Nn/Ln)}$$

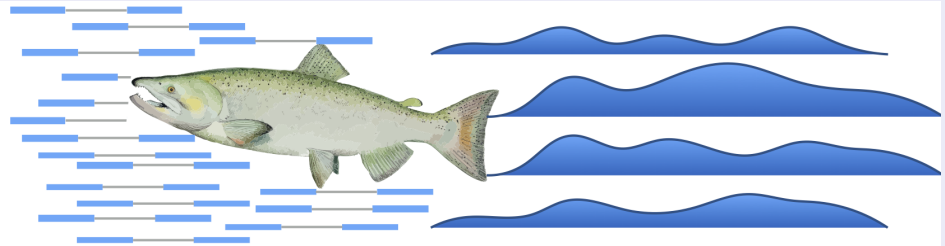
Ni is the number of reads compared to the i-th exon; Li is the length of the i-th exon; sum(N1/L1+N2/L2+.....+Nn/Ln) is the sum of the values of all (n) exons after normalization by length.

- Normalise for gene length first
- And then normalise for sequencing depth and use that to scale number of reads

Salmon is a tool for wicked-fast transcript quantification from RNA-seq data

`https://salmon.readthedocs.io/en/latest/salmon.html`

Salmon can do both mapping and quantification



- Also known as a psuedo-aligner similar tools like Kallisto
- Uses quasi-mapping (an accurate but fast alignments)
- Philosophy is we dont need about accurate alignments
  `Don't count ... quantify`
- Save time and compute resources
- Shows comparable performance to aligners like bowtie or BWA
- Can use bowtie BAM files as output so quassi-mapping is optional.

Salmon: Quantifying in alignment-based mode

We use salmon to count transcripts

`https://salmon.readthedocs.io/en/latest/salmon.html#quantifying-in-alignment-based-mode`

- We use Salmon as the output can be seamlessly imported in DESeq2
- There numerous other alternatives e.g Kallisto, RSEM
- We provide our transcriptome reference and BAM files
- Salmon output file is "quant.sf" with five columns

Name	Name of the target transcript.
Length	length of the target transcript in nucleotides.
EffectiveLength	Estimate based on probability, fragment length distribution biases.
TPM	Estimate of Transcripts Per Million (TPM)
NumReads	Expected number of reads from each transcript given the structure of the uniquely mapping and multi-mapping reads and the relative abundance estimates